免疫克隆选择图划分方法*

刘汉强
(陕西师范大学 计算机科学学院, 西安 710062)

摘 要: 为了解决谱聚类方法中大规模的相似性矩阵的存储和特征分解困难的问题,利用权核 K-均值算法的目标函数和图谱划分准则的等价性,将图谱划分准则作为免疫克隆选择优化算法的亲和度函数,提出一种利用免疫克隆选择优化算法求解图谱划分问题的新方法——免疫克隆选择图划分方法。该方法在免疫克隆选择操作的过程中引入了一个个体修正算子,使得个体以更快的速度向更优的个体进化。此外,在新方法中还引入了流形距离测度来构造相似性矩阵,使得新算法可以有效处理具有复杂结构的数据。采用人工数据集、USPS 手写体数字识别和 UMIST 人脸识别的仿真实验验证了新方法的有效性和鲁棒性。

关键词: 图划分; 谱聚类; 权核 K-均值; 流形相似性测度; 克隆选择
中图分类号: TP75 **文献标志码**: A **文章编号**: 1001-3695(2012)09-3516-05
doi:10.3969/j.issn.1001-3695.2012.09.084

Immune clone selection graph partition algorithm

LIU Han-qiang
(*School of Computer Science, Shaanxi Normal University, Xi'an* 710062, *China*)

Abstract: In order to solve the problem of the storage and eigendecomposition of the similarity matrix in spectral clustering algorithms, this paper proposed a new method utilizing the immune clone selection optimizing algorithm to solve the graph partition. It utilized the equivalence of the graph partitioning and the weighted kernel K-means objectives and adopted the graph partitioning objective as the affinity function. Especially introduced an individual adjustment operator into the immune clone selection optimizing algorithm, which made the individual to evolve in better direction and higher speed. In addition, it introdced a novel distance measure to construct the similarity matrix, namely manifold distance measure, which made the method behave well in data sets with complex structure. The experimental results on six artificial datasets, the USPS handwritten digit datasets and UMIST face datasets show that the novel method is effective and robust.

Key words: graph partition; spectral clustering; weighted kernel K-means; manifold similarity measure; clone selection

利用图谱划分理论来解决聚类问题是许多学者研究的热点方向。图谱划分是一类基于两点间相似关系的方法[1],该类算法与数据点的维数无关,仅与数据点的个数有关。图谱划分方法不对数据的全局结构作假设,而是首先通过统计局部信息来表示两点属于同一类的可能性,然后根据某一图谱划分准则作决策,将所有数据点划分到不同的数据集合中。众所周知,图谱划分准则的求解问题是一个 NP 难的问题。谱聚类是一种常见的解决图谱划分的方法,它利用数据的拉普拉斯矩阵的特征向量进行聚类,获得划分准则在放松了的连续域中的全局最优解。与其他聚类方法相比,谱聚类算法最大的优势是具有识别非高斯分布的能力,非常适合许多实际问题。目前,谱聚类算法已被成功应用于并行计算[2]、VLSI 设计[3]、图像分割[4,5]、数据挖掘[6,7]等领域。

作为一种有效的全局优化搜索进化算法,免疫克隆选择算法[8]在算法实现上兼顾全局搜索和局部搜索,吸取了遗传算法并行搜索优点,通过接种疫苗和计算亲合度,使得算法快速收敛,同时保持一定的多样性,抑制了早熟现象。本文提出了一种利用免疫克隆选择优化算法求解图谱划分问题的新方法——免疫克隆选择图划方法。该方法不需要求解数据的 Laplacian 矩阵特征向量来实现图谱划分,它把图谱划分准则作为免疫克隆选择优化算法的亲和度函数来进行优化。为了加速优化算法的收敛,利用权核 K-均值算法的目标函数和图谱划分准则的等价性[9],引入了一个个体修正算子,使得个体以更快的速度向更优的方向进化。

1 图谱划分理论与权核 K-均值算法

1.1 图谱划分准则

在图谱划分理论[1]中,任意特征空间中的点集均可表示为一个带权无向图 $G=(V,E,S)$,图上的节点 V 即为特征空间中的点,边集 E 表示每两个节点 (i,j) 之间由一条边连接起来,边的权值为 S_{ij},S_{ij} 表示节点 i 和节点 j 的相似程度,称 S 为相似性矩阵。在图 G 中就把聚类问题转变为在图 G 上的图划分问题,即将图 $G=(V,E,S)$ 划分为 k 个互不相交的子集 V_1, $V_2,\cdots,V_k$,划分后保证每个子集 V_i 内的相似程度较高,不同的集合 V_i 和 V_j 之间的相似程度较低。本文采用 $\mathrm{links}(A,B)$ 表示节点集 A 和 B 邻接值的和,即

$$\mathrm{links}(A,B)=\sum_{i\in A,j\in B}S_{ij} \tag{1}$$

收稿日期: 2012-03-11; **修回日期**: 2012-04-26 **基金项目**: 国家自然科学基金资助项目(61102095);陕西省教育厅科研计划资助项目(11JK1008);陕西省自然科学基础研究计划资助项目(2012JQ8045)

作者简介: 刘汉强(1981-),男,山东莱芜人,讲师,主要研究方向为模式识别、图像处理等(liuhq@snnu.edu.cn).

下面介绍几种划分的准则。率切(ratio-cut)准则[3]为

$$\mathrm{RCut}(G)=\min_{V_1,\cdots,V_k}\sum_{c=1}^{k}\frac{\mathrm{links}(V_c,V\backslash V_c)}{|V_c|}$$

率切准则通过引入类规模平衡项来最小化类间相似性。

规范切(normalized-cut)准则[4]为

$$\mathrm{NCut}(G)=\min_{V_1,\cdots,V_k}\sum_{c=1}^{k}\frac{\mathrm{links}(V_c,V\backslash V_c)}{\mathrm{degree}(V_c)}$$

其中:$\mathrm{links}(V_c,V\backslash V_c)=\mathrm{degree}(V_c)-\mathrm{links}(V_c,V_c)$;$\mathrm{degree}(V_c)=\mathrm{links}(V_c,V)$。规范切准则引入容量的概念来规范化类间相关性,从而考虑了相对于类内连接强度的类间连接。

最小最大切(min-max-cut)准则[10]为

$$\mathrm{MMCut}(A,B)=\min_{V_1,\cdots,V_k}\sum_{c=1}^{k}\frac{\mathrm{links}(V_c,V\backslash V_c)}{\mathrm{links}(V_c,V_c)}$$

它同时最小化类间连接强度,最大化类内连接强度。

1.2　权核 K-均值

经典的 K-均值聚类算法只能对样本进行线性划分,当样本线性不可分时,它的聚类性能是很差的。核 K-均值聚类算法[11]将原空间的待分类样本映射到一个高维的特征空间(核空间)中,使得样本变得线性可分(或近似线性可分)。然后在此空间中进行 K-均值聚类。权核 K-均值[12]就是在核 K-均值的基础上的每个节点上引入一个权值,这种推广与图划分问题建立起了强大的联系。

假设样本集合为 $x_1,x_1,\cdots,x_n$,将其分为 k 类:$\pi_1,\pi_1,\cdots,\pi_k$,π_c 表示第 c 类,m_c 为属于 π_c 的点的均值。权核 K-均值的聚类准则是使下面的目标函数最小:

$$D(\{\pi_c\}_{c=1}^{k})=\sum_{c=1}^{k}\sum_{x_i\in\pi_c}w_i\|\varphi(x_i)-m_c\|^2$$

$$m_c=\frac{\sum_{x_i\in\pi_c}w_i\varphi(x_i)}{\sum_{x_i\in\pi_c}w_i}\qquad(2)$$

其中:权值 w_i 是非负的;φ 为核函数,利用这种非线性映射 $\varphi:R^n\to F,x\to\varphi(x)$,将原空间 R^n 的样本 x 映射到更高维的核空间 F 中,使得样本在核空间中变得线性可分或近似线性可分。

1.3　图谱划分准则与权核 K-均值的等价性

Dhillon 等人[9]指出权核 K-均值和图划分准则具有等价性,它们都可以转换为矩阵的迹的形式。关于各种图划分准则与权核 K-均值的等价性证明,读者可以参考文献[9],这里本文只给出权核 K-均值与 normalized-cut 的等价形式。

最小化权核 K-均值目标函数等价于:

$$\min\sum_{c=1}^{k}\sum_{x_i\in\pi_c}w_i\|\varphi(x_i)-m_c\|^2=\max_{Y}\mathrm{trace}(Y^{\mathrm{T}}W^{\frac{1}{2}}KW^{\frac{1}{2}}Y)\qquad(3)$$

其中:W 是由 w_i 构成的对角阵;核矩阵 K 等于 $\Phi^{T}\Phi$,Φ 是由 $\varphi(x_i)$ 构成的矩阵;Y 表示 $n\times k$ 大小的划分矩阵,且是一个正交矩阵($Y^{\mathrm{T}}Y=I_k$)。

Normalized-cut 准则的目标函数最小化可以化简为

$$\min_{V_1,\cdots,V_k}\sum_{c=1}^{k}\frac{\mathrm{links}(V_c,V\backslash V_c)}{\mathrm{degree}(V_c)}=\max_{\tilde{Y}}\mathrm{trace}(\tilde{Y}^{\mathrm{T}}D^{-\frac{1}{2}}SD^{-\frac{1}{2}}\tilde{Y})\qquad(4)$$

其中:S 为相似性矩阵;D 是一个对角矩阵满足 $D_{ii}=\sum_{j=1}^{n}S_{ij}$;$\tilde{Y}$ 是 $n\times k$ 大小的划分矩阵,也是一个正交矩阵。

文献[9]中指出当式(3)中的 W 等于式(4)中的 D 时,式(3)中 Y 就等于式(4)中的 $\tilde{Y}$。如果令式(3)中的 K 等于 $D^{-1}SD^{-1}$,就可得出权核 K-均值的迹最大化问题等价于 normalized-cut 的迹最大化问题,至此就得出了规范切准则与权核 K-均值算法目标函数的等价性。

2　免疫克隆选择图划分算法设计

2.1　相似性矩阵的构造

把谱图划分理论用于数据聚类前的首要工作是构造相似性矩阵。相似性矩阵中的每一个元素表示两两数据点之间相似性,一般采用高斯函数来构造相似性矩阵。设 $X=\{x_1,x_2,\cdots,x_n\}$ 是特征或数据点的集合,计算相似性矩阵的公式为

$$S_{ij}=\exp(-2\sigma^2\|x_i-x_j\|^2)\qquad(5)$$

其中:σ 是高斯函数的尺度参数,采用的距离度量是欧氏距离。从式(5)可以看出,任意两个数据点之间相似性的计算用的都是同样的尺度参数 σ,它起到了放大与收缩两个点之间欧氏距离的作用。

式(5)的使用仅仅考虑了数据的局部一致性,没有考虑数据的全局一致性,即没有保证位于同一流形上的数据点具有较高的相似性。以 Smile 为例(图 1),按照式(5)计算相似性矩阵会导致点 1 和点 2 之间的相似性高于点 1 和点 3,显然这不是本文所期望的。

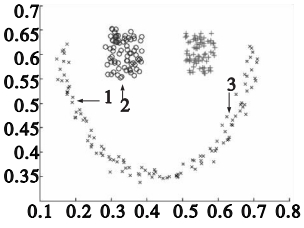

图1　Smile人工数据集

为了在分类问题中既考虑到数据的局部一致性,又充分考虑数据的全局一致性,笔者曾在谱聚类算法中引入了流形距离测度[13]。流形距离测度可以度量沿着流形上的最短路径,使得位于同一流形上的两点可以用许多较短的边相连接,而位于不同流形上的两点要用较长的边相连接,从而实现了放大位于不同流形上的数据点间的距离,而缩短位于同一流形上的数据点间的距离的目的。流形距离测度的具体计算为

$$S_{ij}=\frac{1}{D_{ij}+1}\qquad(6)$$

其中:D_{ij} 表示 x_i 和 x_j 之间的流形距离,即这两个数据点之间的最短路径。具体来说,已知图 G 的顶点集 V,令 P_{ij} 表示连接数据点 x_i 和 x_j 的所有路径的集合,p 表示图上一个长度为 $|p|-1$ 的连接点 p_1 和 $p_{|p|}$ 的路径,此路径经过的边为 (p_h,p_{h+1}),此时流形距离 D_{ij} 为

$$D_{ij}=\min_{p\in P_{ij}}\sum_{h=1}^{|p|-1}L(p_h,p_{h+1})\qquad(7)$$

其中:$L(x_i,x_j)=\rho^{\mathrm{dist}(x_i,x_j)}-1$ 表示两点间流形上的线段长度;$\mathrm{dist}(x_i,x_j)$ 为 x_i 与 x_j 之间的欧氏距离;$\rho>1$ 为伸缩因子。显然,这样定义的线段长度可以用来描述聚类的全局一致性。

2.2　免疫克隆选择图划分

通过流形相似性测度获得相似性矩阵后,需要选择合适的图谱划分准则对数据进行划分,为方便起见,且不失一般性,本文仅以规范切准则为例介绍免疫克隆选择图划分算法,最小化规范切准则可以进一步表示为

$$\min_{V_1,\cdots,V_k}\mathrm{NCut}(G)=\min_{V_1,\cdots,V_k}\sum_{c=1}^{k}\frac{\mathrm{links}(V_c,V\backslash V_c)}{\mathrm{degree}(V_c)}=$$

$$\max_{V_1,\cdots,V_k}\sum_{c=1}^{k}\frac{\mathrm{links}(V_c,V_c)}{\mathrm{degree}(V_c)}=\max_{V_1,\cdots,V_k}\sum_{c=1}^{k}\frac{\sum_{x_i,x_j\in\pi_c}S_{ij}}{\sum_{x_p\in\pi_c}D_{pp}}\qquad(8)$$

在寻优问题中,传统的遗传算法容易陷入局部最小,且收敛速度慢。免疫克隆选择算法是一种新的全局优化搜索算法,其算法实现上兼顾全局搜索和局部搜索,吸取了遗传算法并行搜索优点,通过接种疫苗和计算亲合度,使得算法快速收敛,同时保持一定的多样性,抑制了早熟现象。因此本文采用免疫克隆选择优化算法框架来求解式(8)给出的优化问题。

假设数据集的大小和聚类的数目分别设定为 100 和 5。把免疫克隆选择算法用于实际问题时,首先要解决的就是个体编码问题,本文采用的编码方式为所有数据的类别构成一个个体。个体的变异方法为:首先产生 1 ~ 100 的一个随机整数作为要变异的基因的位置;然后产生一个 0 ~ 1 的随机数,如果该随机数小于变异概率,则该基因位就被变异为 1 ~ 5 的任意整数,且该整数不能和原来基因位上的数相同。本文种群规模设置为 10,克隆比例设置为 5,变异概率设置为 1。

值得指出的是,在本文中每个个体是由各个数据点的类别组成的,这种编码方式使得个体的编码过长,而且产生的个体具有很大的随机性,没有考虑数据点之间的关系,在很多情况下产生的个体可能不具有任何实际意义或与真实分布相背离。更重要的是,优化图谱划分准则又是一个 NP 难的问题。考虑到以上两点,免疫克隆选择算法要获得较好的解,需要很大的迭代次数。

在权核 K-均值算法的框架中,在给定初始划分的基础上,一般根据每个数据点 x_i 与各个聚类中心的距离 $\| \varphi(x_i) - m_c \|^2$ 来产生数据的新划分,实际上这恰恰利用了数据之间的关系来产生新的划分。为了克服本文提出的算法收敛过慢的缺点,受权核 K-均值算法框架的这一处理的启发,定义一个个体修正算子,对随机产生的个体进行修正,使得个体以更快的速度向更优的方向进化。

1.3 节指出了权核 K-均值算法的目标函数和规范切准则具有等价性。在权核 K-均值算法中,$\| \varphi(x_i) - m_c \|^2 = K_{ii} - \frac{2\sum_{x_j \in \pi_c} w_j K_{ij}}{\sum_{x_j \in \pi_c} w_j} + \frac{\sum_{x_j, x_l \in \pi_c} w_i w_j K_{jl}}{(\sum_{x_j \in \pi_c} w_j)^2}$,根据规范切准则和权核 K-均值目标函数的等价性,即当 $W = D, K = D^{-1}SD^{-1}$ 时,可得

$$\| \varphi(x_i) - m_c \|^2 = \frac{S_{ii}}{D_{ii}^2} - \frac{2\sum_{x_j \in \pi_c} S_{ij}}{D_{ii}\sum_{x_j \in \pi_c} D_{jj}} + \frac{\sum_{x_j, x_l \in \pi_c} S_{jl}}{(\sum_{x_j \in \pi_c} D_{jj})^2}$$

式中的第一项对于数据点 x_i 来说是常数,因此该式可以被进一步简化为

$$\| \varphi(x_i) - m_c \|^2 = \frac{\sum_{x_j, x_l \in \pi_c} S_{jl}}{(\sum_{x_j \in \pi_c} D_{jj})^2} - \frac{2\sum_{x_j \in \pi_c} S_{ij}}{D_{ii}\sum_{x_j \in \pi_c} D_{jj}} \quad (9)$$

在对初始种群中的每个个体进行克隆操作之前,先利用式(9)按照数据点 x_i 与各个聚类中心的距离 $\| \varphi(x_i) - m_c \|^2$ 产生一个修正后的新个体,然后计算修正后的个体的亲和度函数值。如果修正后的个体的亲和度函数值大于修正前的,那么就用修正后的个体替代修正前的个体,否则保持不变。对于变异后的个体,在进行选择操作之前,也可进行上述操作,如果修正后的个体的亲和度函数值大于修正前的,那么就用修正后的个体替代修正前的个体,否则保持不变。通过实验发现,这一操作大大加快了算法的收敛速度。

2.3 算法流程及复杂性分析

具体算法描述如下:

a)初始化。选择初始种群 $A(0) = [A_{01}, A_{02}, \cdots, A_{0q}]$,$q$ 为种群的大小。设定最大迭代次数 $Ga = m$。

b)停机判断。判断是否满足终止条件,即是否完成设定的迭代次数。若完成迭代次数,则终止迭代,确定由当前较优个体构成的种群为最优种群,转向步骤 h);否则执行步骤 c)。

c)克隆。对当前的第 g 代父代种群 $A(g)$ 进行克隆操作,得到 $A'(g)$。

d)变异。对 $A'(g)$ 进行变异操作,得到 $A''(g)$。

e)利用式(8)计算亲和度值。计算每个个体亲和度值 Q。

f)克隆选择。若存在变异后的个体 b,使得 $Q(b) = \max(Q(\cdot))$,则选择个体 b 进入新的父代群体。

g)$g = g + 1$,转向步骤 b)。

h)在最优种群中寻找最优个体。按照当前种群中每个个体计算其亲和度值 Q,找到使 Q 取得最大值的个体,确定其为最终结果。

本文算法的复杂度主要是由采用的免疫克隆选择算法的复杂度决定。不失一般性地令免疫克隆选择算法的迭代代数为 Ga,q 为种群规模,克隆比例为 r。本文采用单克隆,主要的算子是克隆变异算子,复杂度为 $O(q)$,因此本文算法的复杂度 $O(Ga * q * r)$。

3 实验比较与结果分析

3.1 多尺度人工数据集

首先将新算法应用于六个人工数据集的聚类问题,六个人工数据集分别为图 1 中的 Smile 和图 2 中的五个数据集,它们具有不同的流形结构,能够用来考察算法对不同结构数据的聚类性能。

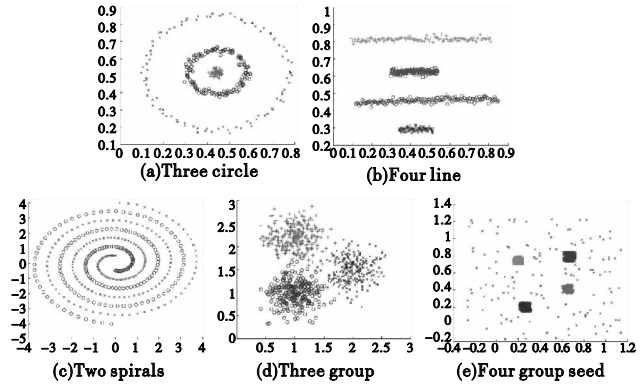

图2 人工数据集

将免疫克隆选择图划分(ICSGP)与原始的 K-均值(KM)算法、谱聚类(SC)算法、基于流形距离的谱聚类[13](MDSC)算法进行性能比较。其中,KM 的最大迭代次数设置为 500,停止阈值设置为 10^{-5};SC(NJW 算法)的参数 $2\sigma^2$ 的变化区间为 $[2^2, 2^{2.2}, \cdots, 2^{10}]$;ICSGP 和 MDSC 的收缩因子 $\rho = e^{ratio}$,ratio 的变化区间也为 $[2^2, 2^{2.2}, \cdots, 2^{10}]$。这样设置参数保证了 ICSGP、MDSC 和 SC 算法具有同样的尺度缩放因子,并确保了各个算法竞争的公平性。此外,本文采用聚类正确率来衡量算法的性能。

本文对每一个数据集独立运行 30 次,各算法在求解以上六个问题时得到的聚类正确率的平均值和最大值如表 1 所示。对于 ICSGP、SC 和 MDSC,先在每一个参数下计算 30 次聚类正确率的平均值和最大值,然后在所有的参数下选取其中的最优值作为算法聚类正确率的平均值和最大值。为了更直观地反映参数对于 ICSGP、MDSC 和 SC 性能的影响,在图 3 中给出了各算法聚类准确率的平均值随参数变化的曲线。

表 1　四种算法在人工数据集上的聚类性能比较

数据集	样本数	维数	类数	ICSGP		MDSC		SC		KM	
				mean	max	mean	max	mean	max	mean	max
smile	266	2	3	1	1	0.9068	1	0.8635	1	0.7652	0.7932
three circle	299	2	3	1	1	0.8844	1	0.9214	1	0.4570	0.4682
four line	512	2	4	1	1	0.8551	1	0.8578	1	0.6520	0.7129
two spirals	378	2	2	1	1	1	1	1	1	0.3819	0.5079
three group	750	2	3	0.9854	0.9867	0.8952	0.9813	0.9246	0.9840	0.9742	0.9867
four group seed	622	2	5	0.9686	0.9727	0.8536	0.9695	0.8435	0.8939	0.7886	0.8215

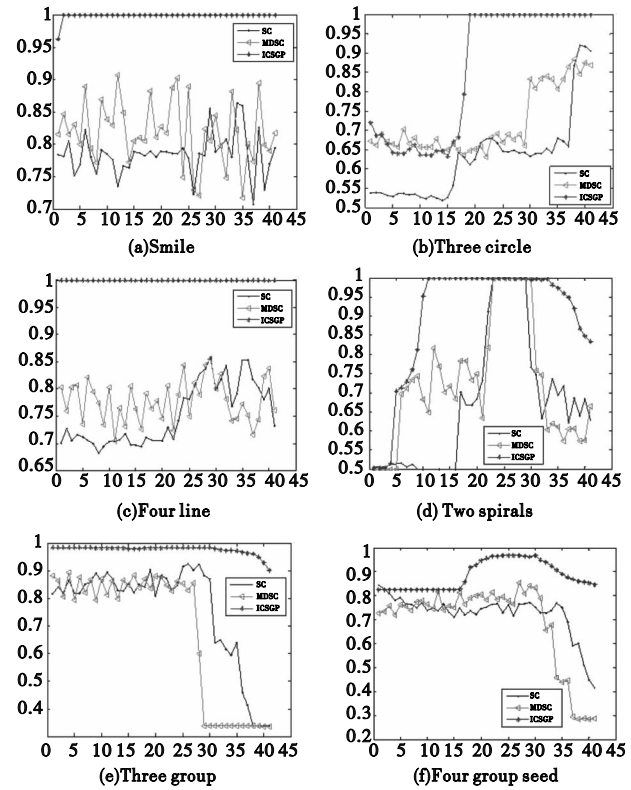

图3　人工数据集上三种方法的聚类准确率随参数变化的曲线

从表 1 可以看出,对流形结构明显的 smile、three circle、four line 和 two spirals 四个问题,ICSGP 算法的聚类正确率的平均值和最大值均达到了 1;对流形结构不明显的 three group 和 four group seed 两个问题,虽然 ICSGP 算法的聚类正确率的平均值和最大值没有达到 1,但它的聚类正确率的平均值和最大值仍是最高的,这充分说明了基于流形距离的相似性度量对复杂结构的数据聚类问题是非常有效的。从表中数据可以看出采用欧氏距离的 SC 算法取得的聚类正确率的最大值与采用流形距离的 MDSC 和 ICSGP 相差无几,这是因为表中数据是所有参数下的最优值,实际上 SC 算法仅能在很少的几个参数下取得好的聚类结果。图 4 给出了对于前四个具有明显流形结构的数据集,ICSGP、MDSC 和 SC 聚类正确率的最大值能达到 1 的参数个数的对比情况。以 smile 为例,MDSC 和 ICSGP 算法的聚类正确率的最高值分别在 41 和 40 个参数下达到了 1,而 SC 算法仅在 8 个参数下达到 1。从图 3 可以看出,对于六个人工数据集,ICSGP 的聚类正确率的平均值在所有参数下几乎都是最高的;对于具有明显流形结构的前四个数据集,MDSC 要优于 SC,对于流形结构不明显的后两个数据集,MDSC 和 SC 不相上下,再次证明了采用流形距离的相似性度量更能反映数据的复杂结构。值得指出的是,在某些数据集下,MDSC 取得的聚类正确率的最大值与 ICSGP 相差无几(如图 4 中的 smile 和 four line),但聚类正确率的平均值却相差很多(如图 3(a)和图 3(c)),这是因为 MDSC 的后续处理中使用了 K-均值聚类算法,其不稳定性导致了 MDSC 的聚类正确率的平均值低于 ICSGP。总体说来,对于人工数据集,ICSGP 的性能最好,MDSC 次之,SC 再次之,KM 最差。

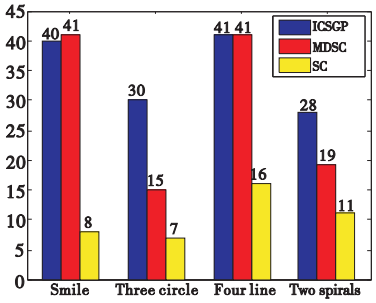

图4　ICSGP、MDSC和SC取得正确分类的参数个数对比

3.2　USPS 手写体数据集算法

本节选择了 USPS 数据集作为测试数据,将新算法应用于手写体数字识别中。USPS 数据集是由 9298 个维灰度图像构成,其中包含 7291 个训练样本,2007 个测试样本。实验取全部测试样本作为聚类数据集,从中挑选三组较难识别的{0,8}、{3,5,8}、{3,8,9}和一组相对容易识别的{1,2,3,4}共四组数字集合进行识别。实验中对于四组数字集合,SC 的参数 $2\sigma^2$ 的变化区间均为$[2^{-9},2^{-8.9},\cdots,2^{-5}]$;ICSGP 和 MDSC 的参数 ratio 的变化区间也作同样的设置;KM 算法的参数设置与 3.1 节相同。对每一个数据集独立运行 30 次,各算法的聚类结果如表 2 所示,表中各数据的统计方法同 3.1 节。在图 5 中给出了 ICSGP、MDSC 和 SC 的聚类准确率的平均值随参数变化的曲线。

表 2　四种算法在 USPS 手写体数据集上的聚类性能比较

手写体集合	样本数	维数	类数	ICSGP		MDSC		SC		KM	
				mean	max	mean	max	mean	max	mean	max
{0,8}	525	256	2	0.9717	0.9810	0.9733	0.9733	0.9429	0.9429	0.7790	0.7886
{3,5,8}	492	256	3	0.8629	0.8780	0.7228	0.8171	0.5381	0.6789	0.5340	0.6585
{3,8,9}	509	256	3	0.9238	0.9312	0.9151	0.9194	0.7682	0.7701	0.4972	0.8153
{1,2,3,4}	828	256	4	0.9453	0.9529	0.9310	0.9324	0.6454	0.7367	0.8155	0.8732

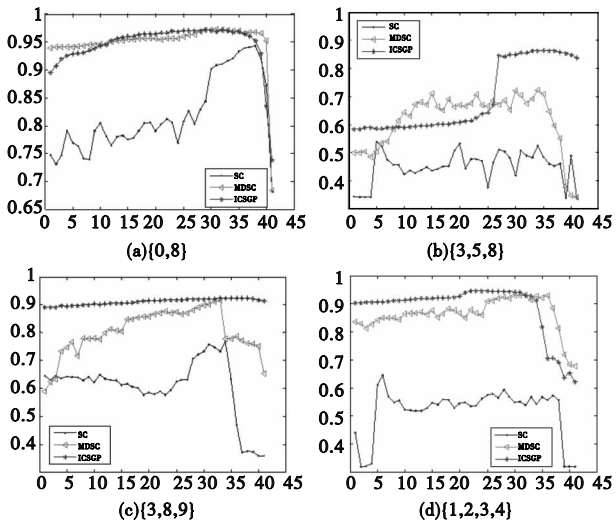

图5　USPS数据集上三种方法的聚类准确率随参数变化的曲线

从表 2 中可以明显看出,无论对三组较难识别的{0,8}、{3,5,8}、{3,8,9}数据集,还是对相对容易识别的{1,2,3,4}数据集,ICSGP 的聚类正确率的平均值和最大值几乎都是四种方法中最优的,MDSC 次之,SC 再次之,KM 最差,表明 ICSGP 在实际应用问题中同样具有良好的性能。从图 5 可以看出,对于 USPS 手写体数据集,ICSGP 和 MDSC 能在比较宽的参数范围内取得理想的聚类结果,SC 在所有参数上的结果均不理想。

3.3 UMIST 人脸数据集

UMIST 人脸库是由 20 个人在相同的光照、不同的姿态(从侧面到正面)条件下,总共 564 张灰度图像组成,图像大小均为 92 × 112。为了下面实验叙述方便,本文对 UMIST_cropped 中的 20 个人按照{1,2,…,20}进行排序。在本文实验中,将人脸图像下采样为 46 × 56 像素大小,从中挑选两组连续的人脸数据集{1;2;3;4;5}和{6;7;8;9;10},以及随机抽取的两组数据集{4;9;12;14;16}和{8;13;14;16;17}作为测试数据。SC 的参数 $2\sigma^2$ 的变化区间为$[2^{-8},2^{-7.9},\cdots,2^{-4}]$;ICSGP 和 MDSC 的参数 ratio 的变化区间也作同样的设置;KM 算法的参数设置同 3.1 节。对每一个数据集独立运行 30 次,各算法的聚类结果如表 3 所示,表中各数据的统计方法同 3.1 节。在图 6 中给出了 ICSGP、MDSC 和 SC 的聚类准确率的平均值随参数变化的曲线。

表 3 四种算法在 UMIST 人脸数据集上的聚类性能比较

UMIST 人脸数据集	样本数	维数	类数	ICSGP		MDSC		SC		KM	
				mean	max	mean	max	mean	max	mean	max
{1;2;3;4;5}	149	5152	5	0.7805	0.9799	0.8174	0.9530	0.7445	0.8255	0.3632	0.6644
{6;7;8;9;10}	116	5152	5	0.8647	0.9569	0.7388	0.9483	0.6486	0.8103	0.4216	0.7155
{4;9;12;14;16}	134	5152	5	0.8209	0.8955	0.7512	0.8284	0.7776	0.8358	0.5565	0.7836
{8;13;14;16;17}	130	5152	5	0.8162	0.9000	0.7513	0.8923	0.6315	0.7462	0.5754	0.7000

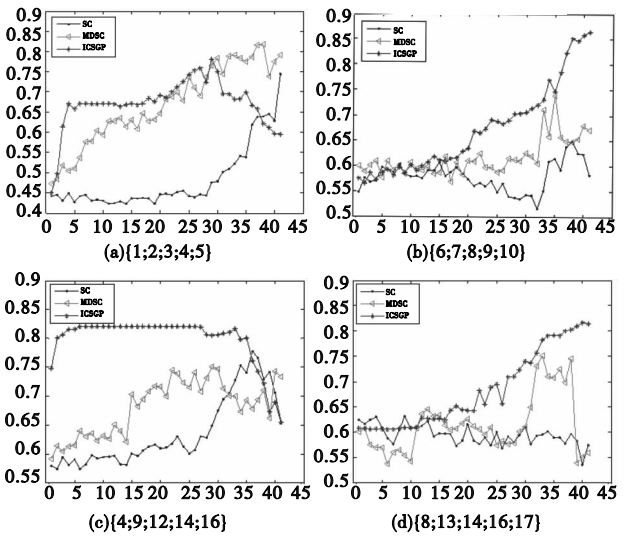

图6 UMIST数据集上三种方法的聚类准确率随参数变化的曲线

从表 3 中可以明显看出,对于人脸图像,ICSGP 的聚类正确率的平均值和最大值几乎都是四种方法中最优的,MDSC 次之,SC 再次之,KM 最差,表明 ICSGP 在人脸识别问题中同样具有良好的性能。从图 6 可以看出,对于 UMIST 人脸数据集,ICSGP 出现了性能不稳定现象,说明人脸图像对于参数比较敏感,但相比而言,ICSGP 的总体性能还是要优于 MDSC 和 SC。

3.4 算法鲁棒性分析

本文采用文献[14]中的鲁棒性分析方法对四种算法在求解以上 14 个问题时的鲁棒性进行比较。具体地,算法 m 在某一特定数据集上的相对性能用该算法获得的聚类正确率的值 C_m 与所有算法在求解该问题时得到的最大的聚类正确率值的比值来衡量,即

$$b_m = \frac{C_m}{\max\limits_k C_k} \tag{10}$$

因此,在某个数据集上表现最好的算法 m^* 的相对性能 $b_{m^*}=1$,而其他算法的相对性能 $b_{m^*}\leqslant 1$。b_m 值越大,则算法在所有算法中的相对性能越好。因此,算法 m 在所有数据集上的 b_m 值的总和可以用来客观评价算法的鲁棒性,总和越大鲁棒性越好。分别针对聚类正确率的最大值和平均值,图 7 给出了四种算法的鲁棒性比较结果,每个算法对应的柱状图顶部所标数值为对应算法在所有 14 个问题上的 b_m 值的总和。

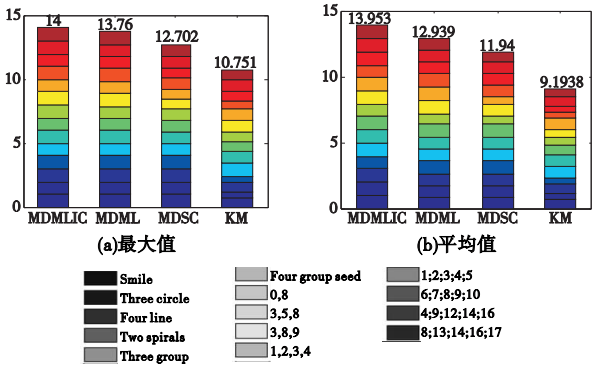

图7 四种算法的鲁棒性比较

从图 7 中可以看出,无论对聚类正确率的最大值还是平均值,ICSGP 均获得了最高的总和值,分别达到了 14 和 13.953;MDSC 也获得了比较满意的值,分别达到了 13.76 和 12.939;而采用欧氏距离作为相似性度量的 SC 和 KM 的总和值均小于采用流形距离作为相似性度量的 ICSGP 和 MDSC。这充分说明了基于流形距离的相似性度量对无监督分类和识别问题具有很好的鲁棒性。ICSGP 聚类正确率的最大值对应的 b_m 值对于测试的 14 个问题均为 1。ICSGP 对不同结构的人工数据聚类、手写体识别以及人脸识别问题均表现出了很好的性能。因此 ICSGP 在所有比较的四种算法中具有最好的鲁棒性。

4 结束语

本文提出了一种新的求解图谱划分问题的方法:免疫克隆选择图划分,可以避免求解数据的 Laplacian 矩阵的特征向量来实现图谱划分。并在新方法中引入了流形距离来构造相似性矩阵,使得新算法可以有效处理具有复杂结构的数据,六个人工数据集、手写体识别以及人脸识别问题的仿真实验表明了新算法的良好聚类性能。值得指出的是,ICSGP 的良好性能是以高的运算时间为代价的。ICSGP 采用免疫克隆算法来优化图谱划分准则,且采用的编码方式造成算法收敛过慢,使得算法的运算时间要明显高于 MDSC、SC 和 KM,因此设计一种更为有效的编码方式将是下一步的研究工作。

参考文献:

[1] FIEDLER M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal,1973,23(98):298-305.

[2] HENDRICKSON B, LELAND R. An improved spectral graph partitioning algorithm for mapping parallel computations[J]. SIAM Journal on Scientific Computing,1995,16(2):452-469.

[3] HAGEN L, KAHNG A B. New spectral methods for ratio cut partitioning and clustering[J]. IEEE Trans Computer-Aided Design, 1992,11(9):1074-1085.

[4] SHI J, MALIK J. Normalized cuts and image segmentation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000,22(8):888-905.

[5] 李俊英,汪西莉. 一种新的大规模复杂图像分割的谱聚类方法[J]. 计算机应用研究,2011,28(5):1994-1997. (下转第 3524 页)

(上接第 3520 页)

[6] 熊忠阳,暴自强,李智星,等. 结合 LSA 的中文谱聚类算法研究[J]. 计算机应用研究,2010,27(3):917-918.

[7] DHILLON I S. Co-clustering documents and words using bipartite spectral graph partitioning[C]//Proc of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2001:269-274.

[8] 焦李成, 杜海峰, 刘芳, 等. 免疫优化计算学习与识别[M]. 北京:科学出版社,2006.

[9] DHILLON I S, GUAN Y, KULIS B. Weighted graph cuts without eigenvectors: a multilevel approach[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2007,29(11):1944-1957.

[10] DING C, HE Xiao-feng, ZHA Hang-yuan, *et al*. A min-max cut algorithm for graph partitioning and data clustering[C]//Proc of ICDM. 2001:107-114.

[11] SCHOLKOPF B, SMOLA A, MULLER K R. Nonlinear component analysis as akernel eigenvalue problem[J]. Neural Computation, 1998,10:1299-1319.

[12] DHILLON I, GUAN Y, KULIS B. Kernel K-means, spectral clustering and normalized cuts[C]//Proc of the 10th ACM KDD Conference. 2004:551-556.

[13] 王玲,薄列峰,焦李成. 密度敏感的谱聚类[J]. 电子学报,2007,35(8):1577-1581.

[14] GENG Xin, ZHAN De-chuan, ZHOU Zhi-hua. Supervised nonlinear dimensionality reduction for visualization and classification[J]. IEEE Trans on Systems, Man, and Cybernetics-Part B, 2005, 35(6):1098-1107.